bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Predictions of Protein-Protein Interactions in *Schistosoma Mansoni*

Javona White Bear,[1,2*] James H. McKerrow[3]

1 Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, University of California, San Francisco, CA 94158

2 Graduate Group in Bioinformatics, University of California, San Francisco, CA 94158, USA

3 Department of Pathology and Sandler Center for Basic Research in Parasitic Diseases, University of California at San Francisco, San Francisco, California 94158, USA

∗*To whom correspondence should be addressed. Email: jwbear15@gmail.com

Abstract

Background

Schistosoma mansoni invasion of the human host involves a variety of cross-species protein-protein interactions. The pathogen expresses a diverse arsenal of proteins that facilitate the breach of physical and biochemical barriers present in skin, evasion of the immune system, and digestion of human hemoglobin, allowing schistosomes to reside in the host for years. However, only a small number of specific interactions between *S. mansoni* and human proteins have been identified. We present and apply a protocol that generates testable predictions of *S. mansoni*-human protein interactions.

Methods

In this study, we first predict *S. mansoni* -human protein interactions based on similarity to known protein complexes. Putative interactions were then scored and assessed using several contextual filters, including the use of annotation automatically derived from literature using a simple natural language processing methodology. Our method predicted 7 out of the 10 previously known cross-species interactions.

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Conclusions

Several predictions that warrant experimental follow-up were presented and discussed, including interactions involving potential vaccine candidate antigens, protease inhibition, and immune evasion. The application framework provides an integrated methodology for investigation of host-pathogen interactions and an extensive source of orthogonal data for experimental analysis. We have made the predictions available online for community perusal.

Author Summary

The *S. mansoni* parasite is the etiological agent of the disease Schistomiasis. However, protein-protein interactions have been experimentally characterized that relate to pathogenesis and establishment of infection. As with many pathogens, the understanding of these interactions is a key component for the development of new vaccines. In this project, we have applied a computational whole-genome comparative approach to aid in the prediction of interactions between *S. mansoni* and human proteins and to identify important proteins involved in infection. The results of applying this method recapitulate several previously characterized interactions, as well as suggest additional ones as potential therapeutic targets.

Introduction

Etiological Agents and Effects of Schistomiasis

Schistosoma are dioecious parasitic trematodes (flukes) that cause the chronic disease Schistomiasis, affecting over 230 million people world-wide and causing more 200,000 deaths a year. They are digenetic organisms with six life cycle stages, four of which take place in the human host. [1] *Schistosoma mansoni*, one of the major etiological agents in Africa and South America of chronic Schistomiasis, releases eggs that become trapped in host tissues, triggering an unsuccessful immune response and causing a host granulomatous reaction, that in its most sever forms, causes fibrosis, scarring, and portal hypertension.

The host granulomatous reaction is a primary cause of mortality associated with Schistomiasis. [2, 3] Infection during childhood frequently results in growth retardation and anemia. The parasite persists in the host for up to 40 years with a high possibility of reinfection in endemic areas. [4] Standard methods for treating Schistosomiasis are not cost effective or affordable in most endemic populations, are ineffective

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

as a prophylaxis against newly acquired infections (i.e. the cercarial and schisotsomula stages of the life cycle), and are becoming increasingly less effective against established infections due to parasite resistance. [5–7] Therefore, there is a need for improved and affordable treatments. [7, 8]

S. mansoni - Human Protein Interactions Involved in Pathogenesis and Infection Across Life Cycle Stages

S. mansoni infection involves parasite- human protein interactions over four of the six parasite life cycle stages [1]. Infection begins during the cercarial stage of the life cycle when the freshwater-dwelling larval cercariae contacts the human host. Invasion of the skin is achieved through degradation of the extracellular matrix by cercarial elastase; this same enzyme may help avoid the host immune response through cleavage of human C3 Complement [9] *S. mansoni* sheds its immunogenic tail to progress to the more mobile schistosomula life cycle stage, where it is carried by blood flow to the lungs and hepatic portal system.

Using proteomic analysis, cercarial elastase was implicated in the cleavage of an extensive list of human proteins, with follow-up experiments confirming its cleavage of at least seven dermal proteins [10]. After schistosomula entry, maturation to the adult life cycle stage occurs in the inferior mesenteric blood vessels where a number of proteins aid in immune evasion and digestion of human hemoglobin proteins. Among the proteins expressed in the adult cycle is the adult tegument surface protein Sm29, a potential Schistosomiasis vaccine candidate antigen. Sm29 interacts with unknown human immune proteins [11]. The final life cycle stage in humans is the egg phase; mated adults produce hundreds of eggs per day to facilitate transmission back to fresh water. The immune reaction to eggs leads to schistome pathogenesis. [12]

Large-Scale Computational Prediction of Protein-Protein Interactions

Ongoing efforts to address Schistomiasis include the development of new vaccines. Knowledge of the specific protein-protein interactions between the pathogen and human host can greatly facilitate this effort. However, a comprehensive literature review revealed only eleven confirmed interactions (Table 4), indicating the characterization effort is still in its infancy. These interactions were identified by experiments such as *in vitro* Edman degradation [10], fluorescence end point assay [13], crystallagrophy [14], and mea-

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

surement of released radioactivity from a suspension [15]. Further types of low-throughput experiments could be based on hypothesis of specific predicted protein interactions [10].

While many methods have been developed to predict intraspecies protein-protein interactions, few have focused specifically on interspecies interactions, where knowledge of the biological context of pathogenesis can be used to refine predictions. Previously, our lab developed a protocol to predict interactions and applied these in the host-pathogen context [16, 17]. Host-pathogen protein complexes were identified using comparative modeling based on a similarity to protein complexes with experimentally determined structures. The binding interfaces of the resulting models were assessed by a residue contact statistical potential, and filtered to retain the pairs known to be expressed in specific pathogen life cycle stages and human tissues (i.e. Biological Context Filter), thus increasing confidence in the predictions. The host-pathogen prediction protocol was benchmarked and applied to predict interactions between human and ten different pathogenic organisms [16, 17].

Informing Computational Predictions

A crucial step for the construction of the Biological Context Filter is to annotate pathogen proteins by the life cycle stages in which they were expressed. This step is especially informative for *S. mansoni*, a digenetic organism, with life cycle stages and protein expression specific to both the molluscum and human hosts. While the human genome has been extensively annotated and made generally available, pathogen genome and expression annotation can be more elusive. Pathogens such as *Plasmodium falciparum* have been sequenced and extensively annotated [18]. In comparison, the *S. mansoni* genome, while much larger than that of *P. falciparum* (11,809 *S. mansoni* proteins vs 5,628 *P. falciparum* proteins), was only recently sequenced and assigned a full set of accession identifiers in GeneDB [19]. The sequencing effort is ongoing and there is limited annotation of corresponding proteins and structural information available. Thus, it is challenging even to cross-reference *S. mansoni* proteins described in various reports, particularly in those published prior to full genome sequencing [20]. Additionally, life cycle stage annotation is difficult to access or even absent in most databases.

The best source of annotation is directly from primary references in literature. Most primary reference accessions were often embedded in portable document format (pdf) files and other file formats that make extraction challenging. Furthermore, the context of an extracted accession and the life cycle stage to which it applies must be isolated from each reference and verifiable for accuracy and further study. We address

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

these challenges by designing a simple natural language processing engine (NLP) that accomplishes data extraction, accuracy, verifiability, and correlation between disperse data sets required to construct the Biological Context Filter.

In Results below, we describe the benchmarking of our host-pathogen prediction protocol against experimentally characterized interactions, followed by a discussion of prospective prediction of novel host-pathogen interactions that may warrant experimental follow-up.

Results & Discussion

Protein Interaction Prediction

The protocol begins with the initial set of 3,052 *S. mansoni* and 8,784 human protein sequences (Fig. 1) for which high-quality models could be created using MODTIE [17]. The initial interaction predictions (Initial Predictions) were obtained by assessing host-pathogen interactions for which comparative models could be constructed. In previous work, the fraction of pathogen proteins that aligned to a protein template of previously observed solved complex structures averaged 21% [17]. Here, only 13.9% of *S. mansoni* proteins could be modeled using such a template. Human proteome interaction template coverage remained consistent with previous work at 34%. Overall, the protocol predicted 528,719 cross-species initial potential interactions. (Table 1).

Next, three network-level filters were applied to prune the initial predictions. The first Network Filter removed interactions based on promiscuous domains, defined as templates used for more than 1% of the total predictions. Promiscuous domains, while present in many interacting complexes lack specificity and are overrepresented in the predicted data set, making them less desirable as vaccine candidate antigens. For example, a domain in the crystal structure of HIV Capsid Protein (p24) bound to FAB13B5, Protein Data Bank (PDB 1E6J), is frequently used template in potential interactions.

However, immunoglobin is conserved in mammals and not specific to human interactions. Many variations will either not be applicable to *S. mansoni* and human interactions or conserved across species for binding similar epitopes. While most of the Fab (fragment antigen binding) region, which form the paratope, is highly variable in sequence, it is composed of less than 22 amino acids and relatively small. Many templates will score above the alignment threshold for this portion of the paratope and the short protein sequence acting as the epitope in the binding site based on shorter sequence lengths, high vari-

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

ability, and conservation across species resulting in a disproportionate number of potential interactions. Templates, similar to p24, were too generalized to draw conclusions with any degree of confidence and removed as promiscuous domains in the first Network Filter.

In the second Network Filter, predictions based on homodimer complexes were removed. This step removes instances where highly conserved interacting dimers form similar complexes in both *S. mansoni* and humans due to speciation events [21, 22]. An example of such is the FGFR2 tyrosine kinase domain (PDB 1E6J). FGFR2 has high similarity to tyrosine kinases in *S. mansoni*, but were generally conserved in eukaryotes and thus comprise a bias in the homodimerization interaction of the catalytic subunits [23]. In the final Network Filter step, interactions with less than a 97% confidence interval were removed to further narrow the focus of potential interactions to a higher confidence level. This filtering results in the remaining 34,164 interactions.

Next, the Biological Context Filter isolates potential interactions according to various life cycle stages of *S. mansoni* and likelihood of in vivo occurrence in human tissues. In the first step of the Biological Context Filter, interactions passing the Network Filter were enriched with data from a simple natural language processing (NLP) literature search and databases using listed nomenclature and functional information (Methods). However, there is limited life cycle stage expression information using database annotation.

An NLP engine was designed to address this limitation and accomplished the following: (1) characterization of 12,720 *S. mansoni* genes automatically from primary reference; (2) recording of contextual, life cycle stage, and citation information into a customized database; and (3) programmatic correlation of this data with existing database annotations. This resulted in annotation of 96.6% of *S.mansoni* sequences, greatly exceeding existing annotation from any single database, which topped out at 62%. NLP annotation further extended this coverage with life cycle stage and characterization information not readily available in database annotation.

Next, the Life Cycle / Tissue Filter refines interactions for likelihood of in vivo interaction based on biological context derived from NLP of the component proteins in each interacting complex and their expression in the four life-cycle stages of *S. mansoni* different human tissues (Materials & Methods). The list of pathogen life cycle stage and human tissue pairs was generated (Table 2).

The third Biological Context Filter applies a targeted post-process analysis of potential interactions. In this step, NLP parameters were used to rank the prediction based on number of occurrences and the

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

assigned weight of the literature where the observation occurred (Materials & Methods).

In previous work predicting host-pathogen protein interactions, filters resulted in a wide range of reductions for different pathogen genome due to varying levels of biological annotation available for each genomes. The majority of the biological annotations in Davis et. al. 2007 [17] were not relevant in a pathogenic context and therefore did not pass the filtering, while pathogen proteins had limited life cycle stage annotation resulting in multiple host-pathogen data sets with no interactions [17].

In the current framework, 22,743 (Table 2) interactions passed both biological and network-level filters, which was 51.5% more than the average of the ten pathogens in the previous work despite a below average model coverage. This increase is largely due to the NLP annotation, which produced a large number of pathogen proteins with a defined life cycle stage. Overall, the Biological Context Filter resulted in 1,345 annotated interactions in likely *in vivo S. mansoni* life cycle stage and human tissue interaction sites (Table 2).

Assessment I: Comparison of predicted and known S. mansoni-human protein interactions.

To assess the predictions, we first compared the predicted set with the set of known *S. mansoni*-human protein interactions. There were 10 confirmed interactions between *S. mansoni* and human proteins. Among the 10, there is only one structure available in PDB (Table 4). The host-pathogen application framework recovered 7 of the 10 known interactions. The majority (7/10) of experimentally characterized *S. mansoni*-human protein interactions involve the serine peptidase cercarial elastase). Several experiments have characterized the cleavage by cercarial elastase of extracellular membrane and complement proteins [10, 12, 24–26].

Our method recapitulated several of these interactions. First, a retrospective prediction was made between the enzyme and human collagen based on the template structure of tick tryptase inhibitor in complex with bovine trypsin (PDB 2UUY) (Fig 2). Previous studies indicate that the enzyme has a role in suppressing host immune response (Table 3) [10, 27]; its similarity to tryptase, which has been used as a indicator of mast cell activations and an important mechanism of host defense against pathogens [28],is consistent with this suggested role of cercarial elastase in pathogenesis.

In addition to cercarial elastases cleavage of extracellular proteins, several important protein-protein interactions involved in *S. mansoni* immune evasion have been characterized, including its cleavage of

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Complement C3 (Table 3). Our method retrospectively predicted this interaction based on the structure of fire ant chymotrypsin in complex with the PMSF inhibitor (PDB 1EQ9) (Fig 2). Fire ant chymotrypsin, which is similar to elastases in many species, degrades proteins for digestion and is a known target for blocking growth from the ant larval stage to adult in ant-infested areas [29].

Assessment II: Comparison of Vaccine Candidate Antigens and S. mansoni-human protein interactions.

Next we assessed predictions against experimentally characterized vaccine candidate antigens where the mechanism, specificity, and interacting human proteins were still undetermined. Currently, there were 9 *S. mansoni* proteins considered as vaccine candidate antigens and 5 protein groups viewed as vaccine candidate antigens. We predicted interactions with 5 of the current vaccine candidate antigens and all of the potential vaccine candidate antigens (Table 5).

We now describe two specific examples of predicted interactions involving *S. mansoni* protein vaccine candidate antigens that may warrant experimental follow-up; these interactions are consistent with literature hypotheses. As noted, cercarial elastase is known to cleave several human proteins (Table 4), and it is considered a vaccine candidate antigen due to its abundance in *S. mansoni* cercarial secretions. Functionally, it has been indicated as the primary means of pathogen entry across the human dermal barriers, the first stage of pathogenesis [24].

We predicted novel interactions between cercarial elastase, its isoforms and other human proteins including calpains, cystatins, tetraspanins, immune and complement proteins (Table 5). An example is the interaction between cercarial elastase and the elastase specific inhibitor elafin.This prediction is based on the template crystal structure of elafin complexed with porcine pancreatic elastase (PDB 1FLE) (Fig 3). Elafin plays a wound-healing role in the dermal immune response in humans and is an antimicrobial against other pathogens such as *Pseudomonas aeruginosa* and *Staphylococcus aureus* [30]. Elafin has been demonstrated to bind with high affinity to both human leukocyte elastase and porcine pancreatic elastase. If a similar affinity can be characterized for its binding to cercarial elastase, elafin could act as a barrier to *S. mansoni*s ability to establish infection. Experimentally validated binding with cercarial elastase could show that increased concentrations of elafin may be sufficient to prevent onset of infection and act as an effective barrier against infection [31].

The Schistosoma protein Sm29 another vaccine candidate antigen indicated in pathogenic immune

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

evasion, was involved in several predictions. The predicted interaction with the human CD59 protein, involved in the complement membrane attack complex (MAC), would aid *S. mansoni*s ability to disable immune response. This prediction was based upon the template of ATF-urokinase and its receptor (PDB 2I9B) (Fig 3), which is involved in multiple patho-physiological processes. Sm29, an uncharacterized transmembrane protein, is an *S. mansoni* surface protein in both the schistosomula and adult life cycle stages that has been indicated in several immune response interactions, making it an important vaccine candidate antigen.

CD59, protectin, regulates complement, inhibits the membrane attack complex (MAC), prevents lysis and is exploited as an established immune evasion tactic used by viruses [32, 33]. Murine experiments indicate immunization with rSm29 reduces *S. mansoni* parasite burdens and offers protective immunity; however, the exact mechanism has not been characterized. Experimental validation of the predicted interaction between Sm29 and CD59 could provide further insight on how *S. mansoni* inhibits the MAC and additional strategies for preventing this inhibition. Additional interactions involving vaccine candidate antigens and key targets are referenced in the supplement.

Limitations

The *S. mansoni* genome was only recently sequenced [20], and there were fewer than 39 validated crystal structures of pathogen proteins available, with only one of these in complex with a human protein. Initial predictions rely on sequence and structure comparison to known interacting complexes, thus the lack of available protein structures in complex limits the coverage of the protocol. Additional experimental efforts will increase coverage and accuracy, by identifying more *S. mansoni* and human protein interactions, more protein interactions in complex, and further characterizing the biology for comparative analysis.

Furthermore, template coverage is primarily restricted to domain- mediated interactions, although peptide-mediated interactions are also known to contribute to protein interaction networks [34]. Peptide motifs that mediate protein interactions were identified through a combination of computational and experimental methods [35,36], and application of these motif-based methods will likely expand the coverage of host-pathogen protein interactions.

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Prediction Errors

Several factors affect the accuracy of the method. These include errors in the comparative modeling process [37], the coarse-grained nature of the statistical potential used to assess the interface residue contacts [16], and consideration of only interactions between individual domains that could lead to predicted interactions that were unfavorable in the context of the full-length proteins. Additionally, both *S. mansoni* and humans are eukaryotic species, which means core cellular components, such as translation machinery, metabolic enzymes, and ubiquitin-signaling components are conserved and comprise many of the initially predicted interactions.

We address the similarities in conserved structures using the Biological Context Filter to remove complexes where there was a low possibility of *in vivo* occurrence, homodimer complexes that clearly involve conserved machinery, and high frequency template domains that could indicate both conserved sequences and structures as well as sequence-structure bias due to lack of interacting template coverage. For example, *S. mansoni* has been shown to secrete chemokine binding proteins as a decoy mechanism that modulates the host immune response. These proteins would be difficult to identify and characterized using known proteomic analysis and would likely be homologous to human proteins and would introduce noise into the detection and isolation of these types of interactions [38].

Future Work

Computational prediction and identification of protein-protein interactions is an important aspect in the development of new vaccines and vaccine candidate antigens. A variety of approaches such as genomic proximity, gene fission/ fusion, phylogenetic tree similarity, gene co-occurrence, co-localization, co-expression, and other features that only make sense or are currently feasible in the context of a single genome [39]. Comparative approaches offer a broad spectrum analysis of protein-protein interactions based on previously observed interactions.

The results of the approach and targeted analysis used on *S. mansoni*-human protein interactions here suggest that sequence and structure-based methods are an applicable approach [16, 40]. This method has several extensions including those that identify peptide motifs [34] , sequence signatures [41] that mediate interactions, and further extension of analysis and interpretation strategies such as analysis of the genetic polymorphisms at loci encoding for the proposed interacting proteins.

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Potential impact

We developed a computational whole-genome method to predict potential host-pathogen protein interactions between *S. mansoni* and humans. Our results show seven validated predictions already experimentally characterized and numerous potential interactions involving proteins indicated as vaccine candidate antigens or potential vaccine candidate antigens. Despite limitations in *S. mansoni* structural coverage, our results demonstrate that broad-spectrum data enrichment and analysis is an effective method for protein-protein interaction prediction and highlight several potential immunization targets against *S. mansoni*. A lists of high-confidence predictions from targeted analysis are available online at http://salilab.org/hostpathogen. Additionally, in the tradition of open source efforts of the biomedical scientific community, the application framework is available for download. In closing, we expect our method to complement experimental methods in and provide insight into the basic biology of *S. mansoni*-human protein interactions.

Materials and Methods

The initial predictions of *S. mansoni*-human were generated based on a protocol described in [17], briefly reviewed here. First, genome-wide *S. mansoni* and human protein structure models were calculated by MODPIPE [42], an automated software pipeline for large-scale protein structure modeling [43]. MODPIPE uses MODELLER [44] to perform the canonical comparative modeling steps of fold assignment, target-template alignment, model construction, and model assessment. High-scoring models were deposited in MODBASE [45], a publicly accessible database of comparative models. Next, resulting models were aligned to SCOP domain sequences, and if a model aligned to a SCOP sequence with more than 70% identity, it was assigned that SCOP domain identifier. These annotations were used as the basis for a search in PIBASE, a database of domain-domain interactions. In this search, those models assigned a SCOP domain that was part of a PIBASE interaction were structurally aligned to the conformation of that domain in the complex. In cases where a human model was aligned to one domain in a PIBASE interaction and an *S. mansoni* model was aligned to the other domain, a putative modeled complex resulted. This complex was then assessed with the MODTIDE potential, which outputs a Z-score approximating the statistical likelihood of the individual domain interface residues forming a complex across the two proteins. A detailed description of the full protocol is available in [16]. We refer to the resulting set of

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

predictions as Initial Predictions.

Filtering Interactions

Two sets of filters were applied to the resulting interactions. The first filter, referred to as the Network Filter was based on aspects of the modeling and scoring process. The second filter, referred to as the Biological Context Filter, was based on the stages of the life cycle and tissue pairs (Figure 1).

Application of Network Filters

Predictions based on templates used for more than 1% of the total number of *S. mansoni* and human interactions were considered promiscuous and removed. 242,677 (45.9%) (Table 1) interactions met this criterion due to the overall similarity in eukaryotic organisms for network level machinery [46] and to the lack of known structure information for *S. mansoni* proteins. High confidence interactions were isolated based on previous work demonstrating an optimal statistical potential Z-score threshold of -1.7, which gave true-positive and false-positive rates of 97% and 3%, respectively [17]. The homodimer complex filter removed predicted interactions based on template complexes formed by protein domains from the same SCOP family excluding highly conserved eukaryotic pathways. These predictions primarily consisted of multimeric enzyme complexes formed by host and pathogen proteins, as well as core cellular components such as ribosome subunits, proteasome subunits, and core cellular components [17]. In total, 143,065 homo-dimer complexes were removed from the filter set based on this criteria (Table 1).

Application of Biological Context Filters

Interactions that pass the Network filtering are then filtered for biological context using the following methods.

Natural Language Processing (NLP) & Enrichment Filters

S. mansoni Protein Annotation.

In preparation for applying the Life Cycle / Tissue Filter, a Natural Language Processing (NLP) protocol

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

was created to automatically identify from the literature which *S. mansoni* proteins were expressed in different pathogen life cycle stages. *S. mansoni* protein database identifiers and their amino acid sequences were extracted from the GeneDB [19], National Center for Biotechnology Information [NCBI], TIGR [47], and Uniprot/TrEMBL databases [48]. A literature search identified experiments indicating proteins expressed in different *S. mansoni* life cycle stages and categorized each literature reference into corresponding life cycle stages. All literature was then mined using NLP to derive accessions and context information. Accessions were derived from the text with regular expression searches corresponding to the specifications of the database (for example, a word in the text matching the regular expression form [A-Z][0-9]5 indicates a Uniprot Accession).

Thus, for each paper, a list of protein accessions was obtained. All protein accessions were then mapped by comparing sequences to Smp accession, Uniprot accession, and NCBI accession, in that order of priority. Thus, the final result of NLP processing was a list of all accessions of proteins expressed in life cycle stages of *S. mansoni*. *S. mansoni* protein sequence data from these initial interactions were enriched from biological annotation obtained from MODBASE [45], GeneDb [19], NCBI, Uniprot [48], and primary reference in literature. The annotations included protein names, links to referenced resources, and any available functional annotation.

Human Protein Annotation.

Human proteins were annotated for tissue expression (GNF Tissue Atlas [49], known expression on cell surface, and known immune system involvement (ENSEMBL [50]. Functional annotation for each protein was obtained from Gene Ontology Annotation (GOA) [51]. Human protein sequences were correlated with predicted interacting sequences to determine involvement [16].

Life Cycle Stage/ Tissue Filter

Next, the Biological Context Filter was applied to *S. mansoni* and human protein interactions in the four life cycle stages associated with pathogenesis and infections in humans. *S. mansoni* proteins were filtered by life cycle stage, known expression and excretion, using NLP and database annotation. An interaction had to be present in the host tissue associated with the specific stage of pathogenesis and that *S. mansoni* life cycle stage to be included in the resulting interactions. The following life cycle stage and tissue pairs were applied to filter interactions: (1) cercariae proteins and human proteins expressed in

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

skin, (2) schistosomula proteins and human proteins expressed in skin, lungs, bronchial, liver, endothelial cells, immune cells, red blood cells, blood, T-cells, early erythroid cells, NK cells, myeloid cells, and B-cells, (3) adult *S. mansoni* and human proteins expressed in liver, endothelial cells, immune cells, red blood cells, blood, T-cells, early erythroid cells, NK cells, myeloid cells, and B-cells, and (4) eggs and human proteins expressed in liver, endothelial cells, immune cells, red blood cells, blood, T-cells, early erythroid cells, NK cells, myeloid cells, and B-cells.

Targeted Filter

The final step in the Biological Context Filter uses a targeted post process analysis based on NLP and database annotations using two additional data mining steps. For each of the interacting protein complex pairs, three parameters were analyzed: pairwise expression in both known human tissue target and *S. mansoni* life cycle stage as indicated by the Life Cycle / Tissue Filter, expression or involvement in known human immunogenic responses, and *S. mansoni* protein expression or involvement with human proteins targeted by other parasites.

Parameters for additional data mining in the target analysis include the following criteria: investigator-selected proteins of interest and NLP derived key terms that were used to target annotation data in protein names and functional annotation (Uniprot [48], GeneDB [19], Gene Ontology [GO] [52]). Proteins selected as targets were assigned weights composed of two factors: (1) an average weight of number of citations across all references to the number of actual references used and (2) an investigator-assigned rank (1-3) based on significance and scope of primary reference/ experiments of NLP sources. The names of proteins and functional annotation were mined for the weighted key terms. All investigator-selected proteins of interest were presumed to pass filter criteria and the remaining interactions were ordered based on key term weights, rank, and Z score.

Assessments

Predictions were benchmarked against confirmed *S. mansoni*-human interactions, which were compiled from the literature. Prospective interactions were assessed using vaccine candidate antigens and hypothesized vaccine candidate antigens where interactions have not been confirmed although several potential human protein binders have been experimentally identified (Table 4). Orthogonal biological information implemented in the filters provided significant enrichment of observed interactions (97% of predicted com-

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

plexes were enriched). The number of protein pairs was reduced by about three orders of magnitude and assessment against previously characterized interactions (63% of known interactions predicted) suggests the method was applicable for genome-wide predictions of protein complexes.

References

1. World Health Organization. WHO _ Schistosomiasis, June 2012. URL `http://www.who.int/topics/schistosomiasis/en/`.

2. ML Burke, MK Jones, GN Gobert, and YS Li. Immunopathogenesis of human schistosomiasis. *Parasite Immunology*, 31(4):163–176, April 2009.

3. Thomas A Wynn, Robert W Thompson, Allen W Cheever, and Margaret M Mentink-Kane. Immunopathogenesis of schistosomiasis. *Immunological reviews*, 201:156–167, October 2004.

4. Center for Disease Control. CDC - Schistosomiasis - Biology, June 2012. URL `http //www.cdc.gov/parasites/schistosomiasis/biology.html`.

5. Gabriela Ribeiro-dos Santos, Sergio Verjovski-Almeida, and Luciana C C Leite. Schistosomiasis—a century searching for chemotherapeutic drugs. *Parasitology Research*, 99(5):505–521, April 2006.

6. Jennifer Keiser Xiao Shuhua Marcel Tanner Burton H Singer Jürg Utzinger. Combination Chemotherapy of Schistosomiasis in Laboratory Studies and Clinical Trials. *Antimicrobial Agents and Chemotherapy*, 47(5):1487, May 2003.

7. Allen G.P. Ross, Paul B. Bartley, Adrian C. Sleigh, G. Richard Olds, Yuesheng Li, Gail M. Williams, and Donald P. McManus. Schistosomiasis — NEJM. *New England Journal of Medicine*, 346(16):1212–1220, 2002.

8. M Ismail, S Botros, A Metwally, S William, A Farghally, L F Tao, T A Day, and J L Bennett. Resistance to praziquantel direct evidence from Schistosoma mansoni isolated from Egyptian villagers. *The American journal of tropical medicine and hygiene*, 60(6):932–935, June 1999.

9. Elizabeth Hansell, Simon Braschi, Katalin F Medzihradszky, Mohammed Sajid, Moumita Debnath, Jessica Ingram, K.C. Lim, and James H McKerrow. Proteomic Analysis of Skin Invasion by Blood Fluke Larvae. *PLoS Neglected Tropical Diseases*, 2(7):e262, July 2008.

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

10. Jessica Ingram, Giselle Knudsen, K.C. Lim, Elizabeth Hansell, Judy Sakanari, and James McKerrow. Proteomic Analysis of Human Skin Treated with Larval Schistosome Peptidases Reveals Distinct Invasion Strategies among Species of Blood Flukes. *PLoS Neglected Tropical Diseases*, 5 (9):e1337, September 2011.

11. Fernanda C Cardoso, Gilson C Macedo, Elisandra Gava, Gregory T Kitten, Vitor L Mati, Alan L de Melo, Marcelo V Caliari, Giulliana T Almeida, Thiago M Venancio, Sergio Verjovski-Almeida, and Sergio C Oliveira. Schistosoma mansoni Tegument Protein Sm29 Is Able to Induce a Th1-Type of Immune Response and Protection against Parasite Infection. *PLoS Neglected Tropical Diseases*, 2(10):e308, October 2008.

12. L.A.L. Quezada, M. Sajid, K.C. Lim, and J.H. McKerrow. A Blood Fluke Serine Protease Inhibitor Regulates an Endogenous Larval Elastase. *Journal of Biological Chemistry*, 287(10):7074–7083, March 2012.

13. Akhmed Aslam, Phyllis Quinn, Richard S McIntosh, Jianguo Shi, Ashfaq Ghumra, James H McKerrow, Karen A Bunting, David W Dunne, Michael J Doenhoff, Sherie L Morrison, Ke Zhang, and Richard J Pleass. Proteases from Schistosoma mansoni cercariae cleave IgE at solvent exposed interdomain regions. *Biochimica et Biophysica Acta (BBA) - Enzymology*, 45(2):567–574, January 2008.

14. Rui Zhao Craig McFarland Jeffrey Kieft Anna Niedzwiecka Marzena Jankowska-Anyszka Janusz Stepinski Edward Darzynkiewicz David N M Jones Richard E Davis Weizhi Liu. Structural Insights into Parasite eIF4E Binding Specificity for m7G and m2,2,7G mRNA Caps. *The Journal of Biological Chemistry*, 284(45):31336, November 2009.

15. S Tzeng, J.H. McKerrow, K Fukuyama, K Jeong, and W L Epstein. Degradation of purified skin keratin by a proteinase secreted from Schistosoma mansoni cercariae. *The Journal of Parasitology*, 69(5):992–994, 1983.

16. F P Davis. Protein complex compositions predicted by structural similarity. *Nucleic Acids Research*, 34(10):2943–2952, May 2006.

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

17. Fred P Davis, David T Barkan, Narayanan Eswar, James H McKerrow, and Andrej Sali. Host-pathogen protein interactions predicted by comparative modeling. *Protein Science*, 16(12):2585–2596, December 2007. URL `http://pibase.janelia.org/modtie/v1.11/index.html`.

18. Malcolm J Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, Karen E Nelson, Sharen Bowman, Ian T Paulsen, Keith James, Jonathan A Eisen, Kim Rutherford, Steven L Salzberg, Alister Craig, Sue Kyes, Man-Suen Chan, Vishvanath Nene, Shamira J Shallom, Bernard Suh, Jeremy Peterson, Sam Angiuoli, Mihaela Pertea, Jonathan Allen, Jeremy Selengut, Daniel Haft, Michael W Mather, Akhil B Vaidya, David M A Martin, Alan H Fairlamb, Martin J Fraunholz, David S Roos, Stuart A Ralph, Geoffrey I McFadden, Leda M Cummings, G Mani Subramanian, Chris Mungall, J Craig Venter, Daniel J Carucci, Stephen L Hoffman, Chris Newbold, Ronald W Davis, Claire M Fraser, and Bart Barrell. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419(6906):498–511, October 2002.

19. F J Logan-Klumpler, N De Silva, U Boehme, M B Rogers, G Velarde, J A McQuillan, T Carver, M Aslett, C Olsen, S Subramanian, I Phan, C Farris, S Mitra, G Ramasamy, H Wang, A Tivey, A Jackson, R Houston, J Parkhill, M Holden, O S Harb, B P Brunk, P J Myler, D Roos, M Carrington, D F Smith, C Hertz-Fowler, and M Berriman. GeneDB–an annotation database for pathogens. *Nucleic Acids Research*, 40(D1):D98–D108, December 2011.

20. Matthew Berriman, Brian J Haas, Philip T LoVerde, R Alan Wilson, Gary P Dillon, Gustavo C Cerqueira, Susan T Mashiyama, Bissan Al-Lazikani, Luiza F Andrade, Peter D Ashton, Martin A Aslett, Daniella C Bartholomeu, Gaelle Blandin, Conor R Caffrey, Avril Coghlan, Richard Coulson, Tim A Day, Art Delcher, Ricardo DeMarco, Appolinaire Djikeng, Tina Eyre, John A Gamble, Elodie Ghedin, Yong Gu, Christiane Hertz-Fowler, Hirohisha Hirai, Yuriko Hirai, Robin Houston, Alasdair Ivens, David A Johnston, Daniela Lacerda, Camila D Macedo, Paul McVeigh, Zemin Ning, Guilherme Oliveira, John P Overington, Julian Parkhill, Mihaela Pertea, Raymond J Pierce, Anna V Protasio, Michael A Quail, Marie-Adèle Rajandream, Jane Rogers, Mohammed Sajid, Steven L Salzberg, Mario Stanke, Adrian R Tivey, Owen White, David L Williams, Jennifer Wortman, Wenjie Wu, Mostafa Zamanian, Adhemar Zerlotini, Claire M Fraser-Liggett, Barclay G

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Barrell, and Najib M El-Sayed. The genome of the blood fluke Schistosoma mansoni. *Nature*, 460 (7253):352–358, July 2009.

21. I Ispolatov. Binding properties and evolution of homodimers in protein-protein interaction networks. *Nucleic Acids Research*, 33(11):3629–3635, June 2005.

22. S K Hanks and T Hunter. Protein kinases 6. The eukaryotic protein kinase superfamily kinase (catalytic) domain structure and classification. *Journal of the Federation of American Societies for Experimental Biology*, 9:576–596, May 1995.

23. Diana Bahia, Luiza Freire Andrade, Fernanda Ludolf, Renato Arruda Mortara, and Guilherme Oliveira. Protein tyrosine kinases in Schistosoma mansoni. *Memórias do Instituto Oswaldo Cruz*, 101:137–143, 2006.

24. P Jones H Sage S Pino-Heiss J H McKerrow. Proteinases from invasive larvae of the trematode parasite Schistosoma mansoni degrade connective-tissue and basement-membrane macromolecules. *Biochemical Journal*, 231(1):47, October 1985.

25. J.H. McKerrow, S Pino-Heiss, R Lindquist, and Z Werb. Purification and characterization of an elastinolytic proteinase secreted by cercariae of Schistosoma mansoni. *Journal of Biological Chemistry*, 260(6):3703–3707, March 1985.

26. Rachel S Curwen, Peter D Ashton, Shobana Sundaralingam, and R Alan Wilson. Identification of novel proteases and immunomodulators in the secretions of schistosome cercariae that facilitate host entry. *Molecular & cellular proteomics MCP*, 5(5):835–844, May 2006.

27. R Correa-Oliveira, E J Pearce, G C Oliveira, D B Golgher, N Katz, L G Bahia, O S Carvalho, G Gazzinelli, and A Sher. The human immune response to defined immunogens of Schistosoma mansoni elevated antibody levels to paramyosin in stool-negative individuals from two endemic areas in Brazil. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 83(6):798–804, November 1989.

28. S He, M D Gaça, and A F Walls. A role for tryptase in the activation of human mast cells: modulation of histamine release by tryptase and inhibitors of tryptase. *The Journal of pharmacology and experimental therapeutics*, 286(1):289–297, July 1998.

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

29. Istvan Botos, Erik Meyer, Myhanh Nguyen, Stanley M Swanson, John M Koomen, David H Russell, and Edgar F Meyer. The structure of an insect chymotrypsin. *Journal of Molecular Biology*, 298 (5):895–901, May 2000.

30. A J Simpson, A I Maxwell, J R W Govan, C Haslett, and J M Sallenave. Elafin (elastase-specific inhibitor) has anti-microbial activity against Gram-positive and Gram-negative respiratory pathogens. *Biochimica et Biophysica Acta (BBA) - Enzymology*, 452(3):309–313, June 1999.

31. O Wiedow, J Lüademann, and B Utecht. Elafin is a potent inhibitor of proteinase 3. *Biochemical and biophysical research communications*, 174(1):6–10, January 1991.

32. Jiusheng Deng, Daniel Gold, Philip T LoVerde, and Zvi Fishelson. Inhibition of the Complement Membrane Attack Complex by Schistosoma mansoni Paramyosin. *INFECTION AND IMMUNITY*, 71(11):6402–6410, November 2003.

33. Zvi Fishelson. Novel mechanisms of immune evasion by Schistosoma mansoni. *Memórias do Instituto Oswaldo Cruz*, 90(2):289–292, 1995.

34. Victor Neduva and Robert B Russell. Peptides mediating interaction networks: new leads at last. *Biochimica et Biophysica Acta (BBA) - Enzymology*, 17(5):465–471, October 2006.

35. Amy Hin Yan Tong, Becky Drees, Giuliano Nardelli, Gary D Bader, Barbara Brannetti, Luisa Castagnoli, Marie Evangelista, Silvia Ferracuti, Bryce Nelson, Serena Paoluzi, Michele Quondam, Adriana Zucconi, Christopher W V Hogue, Stanley Fields, Charles Boone, and Gianni Cesareni. A Combined Experimental and Computational Strategy to Define Protein Interaction Networks for Peptide Recognition Modules. *Science Signaling*, 295(5553):321, January 2002.

36. Victor Neduva, Rune Linding, Isabelle Su-Angrand, Alexander Stark, Federico de Masi, Toby J Gibson, Joe Lewis, Luis Serrano, and Robert B Russell. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS biology*, 3(12):e405, December 2005.

37. M A Marti-Renom, A C Stuart, A Fiser, R Sánchez, F Melo, and A Sali. Comparative protein structure modeling of genes and genomes. *Annual review of biophysics and biomolecular structure*, 29(1):291–325, 2000.

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

38. Philip Smith, Rosie E Fallon, Niamh E Mangan, Caitriona M Walsh, Margarida Saraiva, Jon R Sayers, Andrew N J McKenzie, Antonio Alcami, and Padraic G Fallon. Schistosoma mansoni secretes a chemokine binding protein with antiinflammatory activity. *Journal of Experimental Medicine*, 202(10):1319–1325, November 2005.

39. Benjamin A Shoemaker and Anna R Panchenko. Deciphering Protein–Protein Interactions. Part II. Computational Methods to Predict Protein and Domain Interaction Partners. *PLOS Computational Biology*, 3(4):e43, April 2007.

40. Haiyuan Yu, Nicholas M Luscombe, Hao Xin Lu, Xiaowei Zhu, Yu Xia, Jing-Dong J Han, Nicolas Bertin, Sambath Chung, Marc Vidal, and Mark Gerstein. Annotation Transfer Between Genomes: Protein–Protein Interologs and Protein–DNA Regulogs. *Genome Research 2004 14: 1107-1118*, 14:1107–1118, 2004.

41. E Sprinzak and H Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*, 311(4):681–692, 2001.

42. U Pieper, B M Webb, D T Barkan, D Schneidman-Duhovny, A Schlessinger, H Braberg, Z Yang, E C Meng, E F Pettersen, C C Huang, R S Datta, P Sampathkumar, M S Madhusudhan, K Sjolander, T E Ferrin, S K Burley, and A Sali. ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Research*, 39(Database):D465–D474, December 2010.

43. Narayanan Eswar, Bino John, Nebojsa Mirkovic, Andras Fiser, Valentin A Ilyin, Ursula Pieper, Ashley C Stuart, Marc A Marti-Renom, M S Madhusudhan, Bozidar Yerkovich, and Andrej Sali. Tools for comparative protein structure modeling and analysis. *Nucleic Acids Research*, 31(13): 3375–3380, 2003.

44. Andrej Sali and Tom L Blundell. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Biochimica et Biophysica Acta (BBA) - Enzymology*, 234(3):779–815, December 1993.

45. M Shen and A Sali. Statistical potential for assessment and prediction of protein structures. *Protein Science*, 15(11):2507–2524, 2009.

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

46. M K Basu, L Carmel, I B Rogozin, and E V Koonin. Evolution of protein domain promiscuity in eukaryotes. *Genome research*, 18(3):449–461, 2008.

47. Feng Liang Ingeborg Holt Geo Pertea Jonathan Upton John Quackenbush. The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Research*, 28(1):141, January 2000.

48. The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 40(D1):D71–D75, December 2011.

49. Andrew I Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, Michael P Cooke, John R Walker, and John B Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences*, 101(16):6062–6067, April 2004.

50. T J P Hubbard, B L Aken, K Beal, B Ballester, M Caccamo, Y Chen, L Clarke, G Coates, F Cunningham, T Cutts, T Down, S C Dyer, S Fitzgerald, J Fernandez-Banet, S Graf, S Haider, M Hammond, J Herrero, R Holland, K Howe, K Howe, N Johnson, A Kahari, D Keefe, F Kokocinski, E Kulesha, D Lawson, I Longden, C Melsopp, K Megy, P Meidl, B Ouverdin, A Parker, A Prlic, S Rice, D Rios, M Schuster, I Sealy, J Severin, G Slater, D Smedley, G Spudich, S Trevanion, A Vilella, J Vogel, S White, M Wood, T Cox, V Curwen, R Durbin, X M Fernandez-Suarez, P Flicek, A Kasprzyk, G Proctor, S Searle, J Smith, A Ureta-Vidal, and E Birney. Ensembl 2007. *Nucleic Acids Research, Database issue*, 35:D610–D617, 2007.

51. Evelyn Camon, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, John Maslen, David Binns, Nicola Harte, Rodrigo Lopez, and Rolf Apweiler. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research, Database issue*, 32:D262–D266, 2004.

52. David Botstein, J Michael Cherry, Michael Ashburner, Catherine A Ball, Judith A Blake, Heather Butler, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, Midori A Harris, David P Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C Matese, Joel E Richardson, Martin Ringwald, Gerald M Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology - Nature Genetics. *Nature*, 25(1):25–29, May 2000.

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

53. Marc H Dresden and Harold L Asch. Proteolytic enzymes in extracts of Schistosoma mansoni Cercariae. *Biochimica et Biophysica Acta (BBA) - Enzymology*, 289(2):378–384, December 1972.

54. J P Salter. Schistosome Invasion of Human Skin and Degradation of Dermal Elastin Are Mediated by a Single Serine Protease. *Journal of Biological Chemistry*, 275(49):38667–38673, September 2000.

55. G. Gazzinelli and J. Pellegrino. Elastolytic Activity of Schistosoma Mansoni Cercarial Extract. *Journal of Parasitology*, 50:591–592, August 1964.

56. William Castro-Borges, Adam Dowle, Rachel S Curwen, Jane Thomas-Oates, and R Alan Wilson. Enzymatic Shaving of the Tegument Surface of Live Schistosomes for Proteomic Analysis A Rational Approach to Select Vaccine Candidates. *PLoS Neglected Tropical Diseases*, 5(3):e993, March 2011.

57. A A Da'dara, P J Skelly, M M Wang, and D A Harn. Immunization with plasmid DNA encoding the integral membrane protein, Sm23, elicits a protective immune response against schistosome infection in mice. *Vaccine*, 20(3-4):359–369, November 2001.

58. Giselle M Knudsen, Katalin F Medzihradszky, Kee-Chong Lim, Elizabeth Hansell, and James H McKerrow. Proteomic analysis of Schistosoma mansoni cercarial secretions. *Molecular & cellular proteomics MCP*, 4(12):1862–1875, December 2005.

59. Melaine Delcroix, Katalin Medzihradsky, Conor R Caffrey, Richard D Fetter, and James H McKerrow. Proteomic analysis of adult S. mansoni gut contents. *Molecular & Biochemical Parasitology*, 154(1):95–97, July 2007.

60. M Tendler, C A Brito, M M Vilar, N Serra-Freire, C M Diogo, M S Almeida, A C Delbem, J F Da Silva, W Savino, R C Garratt, N Katz, and A S Simpson. A Schistosoma mansoni fatty acid-binding protein, Sm14, is the potential basis of a dual-purpose anti-helminth vaccine. *Proceedings of the National Academy of Sciences of the United States of America*, 93(1):269–273, January 1996.

61. Maged Al-Sherbiny, Ahmed Osman, Rashida Barakat, Hala El Morshedy, Robert Bergquist, and Richard Olds. In vitro cellular and humoral responses to Schistosoma mansoni vaccine candidate antigens. *Acta tropica*, 88(2):117–130, October 2003.

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

62. Cristina T Fonseca, Edécio Cunha-Neto, Anna C Goldberg, Jorge Kalil, Amélia R de Jesus, Edgard M Carvalho, Rodrigo Correa-Oliveira, and Sergio C Oliveira. Human T cell epitope mapping of the Schistosoma mansoni 14-kDa fatty acid-binding protein using cells from patients living in areas endemic for schistosomiasis. *Microbes and infection / Institut Pasteur*, 7(2):204–212, February 2005.

63. E J Pearce, S L James, S Hieny, D E Lanar, and A Sher. Induction of protective immunity against Schistosoma mansoni by vaccination with schistosome paramyosin (Sm97), a nonsurface parasite antigen. *Proceedings of the National Academy of Sciences of the United States of America*, 85(15): 5678–5682, August 1988.

64. Kamal A Shalaby, Lei Yin, Arvind Thakur, Linda Christen, Edward G Niles, and Philip T LoVerde. Protection against Schistosoma mansoni utilizing DNA vaccination with genes encoding Cu/Zn cytosolic superoxide dismutase, signal peptide-containing superoxide dismutase and glutathione peroxidase enzymes. *Vaccine*, 22(1):130–136, December 2003.

65. Afzal A Siddiqui, Troy Phillips, Hugues Charest, Ron B Podesta, Martha L Quinlin, Justin R Pinkston, Jenny D Lloyd, Janet Pompa, Rachael M Villalovos, and Michelle Paz. Enhancement of Sm-p80 (large subunit of calpain) induced protective immunity against Schistosoma mansoni through co-delivery of interleukin-2 and interleukin-12 in a DNA vaccine formulation. *Vaccine*, 21 (21-22):2882–2889, June 2003.

66. Sophia J Parker-Manuel, Alasdair C Ivens, Gary P Dillon, and R Alan Wilson. Gene Expression Patterns in Larval Schistosoma mansoni Associated with Infection of the Mammalian Host. *PLoS Neglected Tropical Diseases*, 5(8):e1274, August 2011.

67. M H Tran, M S Pearson, J M Bethony, and D J Smyth. Tetraspanins on the surface of Schistosoma mansoni are protective antigens against schistosomiasis. *Nature medicine*, 2006.

68. Andreas Ruppel, Diane J McLaren, Hans Jochen Diesfeld, and Ursula Rother. Schistosoma mansoni escape from complement-mediated parasiticidal mechanisms following percutaneous primary infection. *European Journal of Immunology*, 14(8):702–708, 1984.

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure Legends

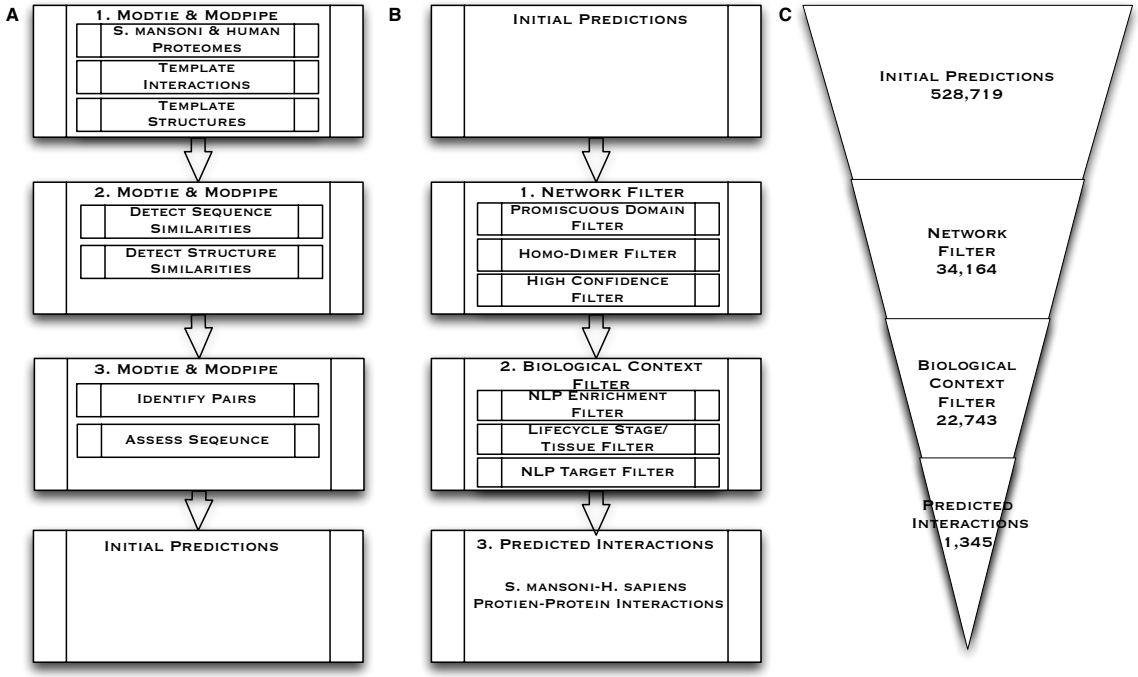

Figure 1. Prediction Framework: (A) Modtie & ModPipe protocol for detecting sequence and structure similarity: 1. The protocol begins with the set of human and *S. mansoni* proteins. 2. Sequence matching procedures are then used to identify similarities between the proteins and proteins with known structure or interactors. 3. A structure-based statistical potential assessment, or a sequence similarity score in the absence of structure, is then used to identify pairs with similarity to known complexes, assess the basis for a putative interaction, predict interacting partners and yield the Initial Predictions. (B) Initial Predictions: 1. Network Filter: Promiscuous Dimers, homo-dimer complexes, and high confidence interactions are then extracted from the initial set of predictions. 2. The Biological Context Filter is applied to the remaining set of predictions, weighing and ranking NLP enriched predictions, isolating life cycle stage and tissues interactions between *S. mansoni* and humans, and application of the Targeted Filter. 3. Predicted Interactions are an output of the Biological Context Filter. (B) Illustrates the reduction in interactions obtained after each step. The framework reduces the number of potential *S. mansoni*-human protein interactions by about three orders of magnitude (Table 1-3).

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

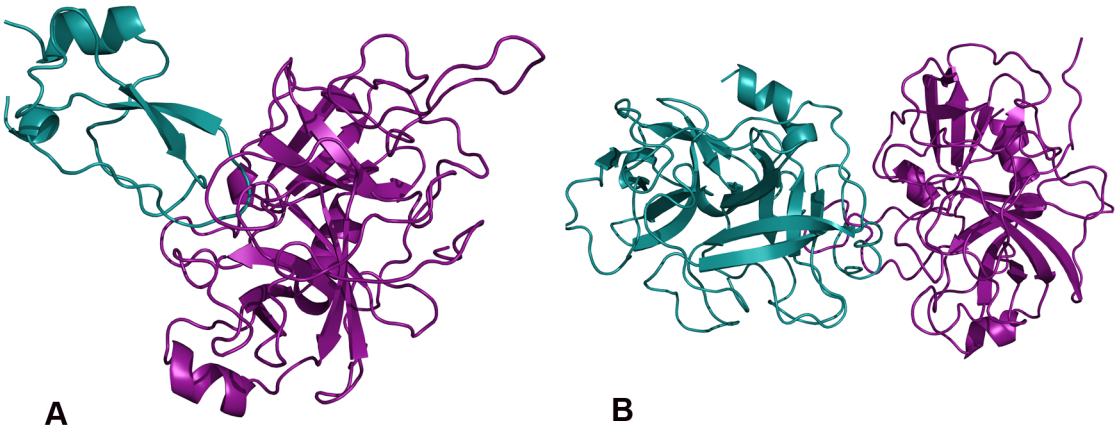

Figure 2. Retrospective Predictions: Examples of validated interactions. (A) Cercarial elastase (purple) and human collagen (blue) based on the template structure of tick tryptase inhibitor in complex with bovine trypsin (PDB 2UUY) (B) Cercarial elastase (purple) and human Complement C3 (precursor C3b) (blue) based on the template structure (PDB 1EQ9) of fire ant chymotrypsin complexed with PMSF, an inhibitor. Figures were generated by PyMOL (http://www.pymol.org).

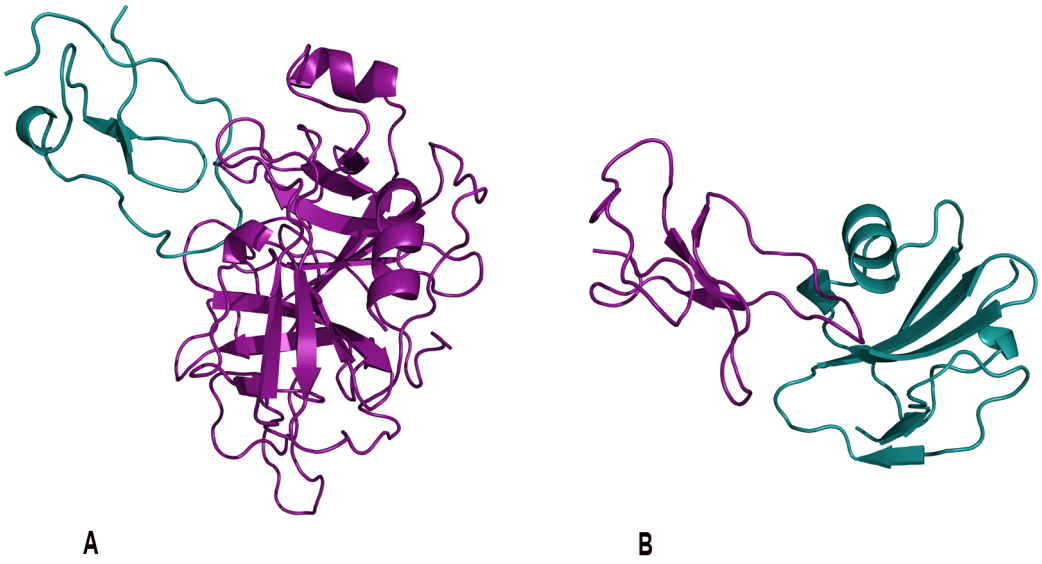

Figure 3. Prospective Predictions: Examples of predicted interactions. (A) Cercarial elastase (purple) was predicted to interact with the human elastase specific inhibitor elafin (blue). This prediction is based on the template crystal structure of elafin complexed with porcine pancreatic elastase (PDB 1FLE). (B) Sm29 (purple) was predicted to interact with human CD59 (blue) protein involved in the membrane attack complex corroborates hypothesis of *S. mansoni*'s ability to disable immune response. This prediction was based upon the template of ATF-urokinase and its receptor (PDB 2I9B). Sm29, an uncharacterized transmembrane protein, is an *S. mansoni* surface protein in both the schistosomula and adult life cycle stages that has been indicated in several immune response interactions making it an important vaccine candidate antigen. Figures were generated by PyMOL (http://www.pymol.org).

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Tables

Network Filter.		
Filter	**Number of Interactions Removed**	**Interactions Remaining**
Unfiltered	-	528,719
Promiscuous Domains	242,677	286,042
Homo-dimer Complexes	143,065	142,977
High Confidence	108,813	34,164

Table 1. Network Filter: Interactions removed during each application of indicated filter. Unfiltered interactions from the initial result set are shown in the first row. The number of interactions removed and remaining from each filtering step are shown in each category column including Promiscuous Pairs, Homo-dimer Complexes, and High Confidence interactions. The remaining 34,164 interactions are used as an input into the next step, the Biological Context Filter.

Biological Context Filter Interactions.		
Filter	**Interactions Removed**	**Interactions Remaining**
Network	-	34,164
NLP & Enrichment	10,929	23,235
Life cycle/ Tissue	492	22,743
Targeted	21,398	1,345

Table 2. Biological Context Filter: Interactions removed during each application of indicated filter. Unfiltered interactions from the Network Filter result set are shown in the first row. The number of interactions found from the total resulting data set are shown in each category column.

Biological Context Filter Interactions and *S. mansoni* Life Cycle Stage.		
Life cycle stage	**Correlated Interactions**	**Life cycle stage-Tissue**
Cercariae	460 (460/1345)	1 (skin)
Schistosomula	442 (442/1345)	15
Adult	329 (329/1345)	13
Egg	114 (114/1345)	11

Table 3. Protein interactions involved in *S. mansoni* life cycle stages that are directly involved in human pathogenesis from the resulting targeted predictions are shown with targeted life cycle stage, the number of predicted interactions for the corresponding life cycle stage and associated human tissues involved in the interaction. Human tissue expression data were obtained from the GNF Tissue Atlas [49] and GO [52] functional annotation unless noted otherwise.

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Comparison of known and predicted *S. mansoni* protein interactions.				
S. Mansoni Protein	**Human Protein**	**Predicted**	**Reference**	**PDB**
Smp_001500 EIF4E	EIF4E-binding protein 1	No	[14]	3HXG
SmCe Cercarial elastase	Collagen (I, IV, VIII)	Yes	[15, 24, 25]	2CHA
SmCe Cercarial elastase	IgE	No	[13]	-
SmCe Cercarial elastase	Complement C3 (C3b)	Yes	[10]	1EQ9
SmCe Cercarial elastase	Laminin	Yes	[15, 24, 25]	2CHA
SmCe Cercarial elastase	Fibronectin	Yes	[15, 24, 25]	2CHA
SmCe Cercarial elastase	Keratin	Yes	[53]	-
SmCe Cercarial elastase	Elastin	Yes	[54, 55]	1FON
SmCB2 Cathepsin B	Collagen (I) (nidogen)	Yes	[10]	1STF
SmCB2 Cathepsin B	Complement C3	No	[10]	-

Table 4. Confirmed protein-protein interaction between S. mansoni and human proteins. The application framework predicted 7 of the 10 known interactions. PDB column indicates the template PDB structure used to predict the interaction.

bioRxiv preprint doi: https://doi.org/10.1101/233072; this version posted December 28, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Potential *S. mansoni* Protein Interactions.				
***S. Mansoni* Protein**	**Human Protein**	**Predicted**	**Reference**	**PDB**
Sm-TSP-1 Tetraspanin	IgG1/IgG3 Immune Response	No	[56]	-
Sm-TSP-2 Tetraspanin	IgG1/IgG3 Immune Response	No	[56]	-
Sm 29 Transmembrane	67782326 (GI) TGF-beta receptor	Yes	[11, 56]	2I9B, 1YWH
Sm 29 Transmembrane	67782324 (GI) TGF, beta receptor II	Yes	[11, 56]	2I9B, 1YWH
Sm 29 Transmembrane	42716302 (GI) CD59 glycoprotein precursor, MAC	Yes	[11, 56]	2I9B, 1YWH
Sm 29 Transmembrane	9966907 (GI) SLURP-1	Yes	[11, 56]	2I9B, 1YWH
Sm 29 Transmembrane	4505865 (GI) PLAU	Yes	[11, 56]	2I9B, 1YWH
Sm 29 Transmembrane	53829381 (GI) PLAU	Yes	[11, 56]	2I9B, 1YWH
Sm 29 Transmembrane	53829379 (GI) PLAU	Yes	[11, 56]	2I9B, 1YWH
Sm 29 Transmembrane	4504033 (GI) GPI anchored molecule-like	Yes	[11, 56]	2I9B, 1YWH
Sm 23 Tetraspanin	IgG3, MAP-3 Immune Response	No	[11, 56–58]	-
Sm 14 FABP	IgG1, IgG3 Immune Response	No	[11, 26, 56, 58–62]	-
Sm 97 Paramyosin	IgG, IgE Immune Response	Yes	[58, 61, 63]	-
Sm 28 GST	IL-5, IgG2, MAP-4 Immune Response	No	[11, 58, 61]	-
SOD SOD [Cu-Zn], Cytosolic	4507149 (GI) SOD1 [Cu-Zn]	Yes	[11, 58, 59, 64]	2AF2, 1JK9
SOD SOD [Cu-Zn], Cytosolic	118582275 (GI) SOD3 Extracellular	Yes	[11, 58, 59, 64]	2AF2
SOD SOD [Cu-Zn], Cytosolic	4826665 (GI) Copper chaperone for SOD	Yes	[11, 58, 59, 64]	2AF2, 1JK9
Sm-p80 Katanin p80 WD40	C3 Complement Immune Response	No	[56, 65]	-
Cercarial Elastase	4505787 (GI) Elafin *Supplemental Table 1	Yes	[31]	1FLE
Venom Allergen Proteins (VAL)	*Supplemental Table 2	Yes	[26, 56, 66]	*
Calpain	*Supplemental Table 3	Yes	[11, 12, 58, 65, 66]	*
Cystatin	*Supplemental Table 4	Yes	[58, 59]	*
Tetraspanin	*Supplemental Table 5	Yes	[11, 56, 66, 67]	*
Immune evasion	Immunoglobin Proteins *Supplemental Table 6	Yes	[5, 11, 13, 32, 59, 68]	*

Table 5. Prospective protein-protein interactions between *S. mansoni* and human proteins. Prospective interactions are hypothesized, but have little or no experimental evidence, and are currently under investigation as candidate antigens or potential vaccine candidate antigens. The application framework predicted interactions between several proteins and suggested *S. mansoni* and human interactions with further interactions listed in Supplemental Tables (1-6).